基于自动机器学习流程优化的雷达辐射源信号识别 *

涂同珩，金炜东

(西南交通大学 电气工程学院，成都 610031)

摘　要：针对雷达辐射源信号识别课题中复杂的参数配置问题，从机器学习参数优化的研究入手，发现了一种基于树结构的机器学习流程优化方法，该方法利用遗传编程生成基于树结构的机器学习流程，并对其结构和参数进行进化，得到表现最佳的带参数的机器学习流程。该流程可以包括特征处理和建模的任意组合，实现对原始数据集的学习和识别。并与人工参数配制的一对一支持向量机在两种不同维度的雷达信号特征集上进行对比识别，相比之下，该方法无须繁琐的参数配置，最高准确率提高超过 6%，证明该方法得到的基于树结构的机器学习流程有着明显的优势。

关键词：自动机器学习；超参数优化；遗传编程；雷达辐射源信号；支持向量机

中图分类号：TN911.6　　doi: 10.3969/j.issn.1001-3695.2017.07.0686

Radar emitter signal recognition based on optimization of automatic machine learning pipeline

Tu Tongheng, Jin Weidong

(*School of Electrical Engineering*, *Southwest Jiaotong University*, *Chengdu* 610031, *China*)

Abstract: In order to solve the problem of complex parameter configuration in radar emitter signal recognition, this paper found a tree-based machine learning pipeline optimization method, based on the study of machine learning parameter optimization. This method uses GP to generate the tree-based machine learning pipeline, and gets the best machine learning pipeline with the best parameters through the evolution of the pipeline and parameters. The pipeline could include any combination of feature processing and modeling to achieve the learning and recognition of the original data set. This paper compared recognition effect of this method and the SVM method in the radar emitter signal feature sets of two different dimensions. In contrast, this method does not require cumbersome parameter configuration, and the highest accuracy is improved by more than 6%. It is proved that the tree-based machine learning pipeline obtained by this method has obvious advantages.

Key Words: auto ML; hyperparameter optimization; genetic programming; radar emitter signal; SVM

0　引言

由于以相控阵雷达为代表的新体制雷达的投入使用与战场中辐射源数量的急剧增加，在现代战场电磁环境下，信号越来越密集，信号样式越来越复杂，这对雷达辐射源信号的识别带来越来越大的挑战[1]。现阶段的识别研究基本围绕特征提取、特征选择和分类器分类这三个部分展开，识别的方法主要以机器学习为主[2]。

目前，在雷达辐射源信号识别的研究中，使用的信号特征有直接测量得到的到达时间(TOA)、载频(RF)、脉宽(PW)、脉幅(PA)、到达方向(DOA)等，也包括越来越受到重视的脉内特征[2]。主要采用的机器学习方法有支持向量机(SVM)、人工神经网络(ANN)、随机森林(RF)等[3]。

由于基本采用机器学习的方法，雷达辐射源信号的识别问题终究也属于机器学习问题的一部分，当前典型的机器学习流程是先对原始数据进行数据清洗，经过特征提取、特征构建、特征选择、最后交由经过参数优化的分类模型进行分类。在这套流程中，实验者需要以某种方式转换数据，使其更适合于建模，例如通过对特征进行归一化（即特征变换），去除对于建模不太有用的特征（即特征选择），或从现有数据创建新特征（即特征构造）。然后实验者必须选择合适的机器学习模型（即模型选择），并选择使分类结果最准确的模型参数（即参数优化），以保证最后的分类准确度。有经验的专家尚难以较为轻松地设置好这些步骤，更不论在信号日益复杂的背景下其他人员配置步骤和参数的能力了[4,5]。

在过去的 20 年中已经看到了智能系统突飞猛进的发展，它们在如太空天线的设计、大型软件项目漏洞的发现和修补，甚至与人围棋对弈等各种各样的任务中，都能够有明显超越人类的表现。可见这些智能系统有着巨大的创造力，那么，智能系统可以自动设计机器学习流程吗？

基金项目：国家自然科学基金重点资助项目（61134002）；中央高校基本科研业务费专项资金资助项目（SWJT12CX038U）

作者简介：涂同珩（1993-），男，湖北武人，硕士研究生，主要研究方向为信号处理与模式识别（totoheng@sina.com）；金炜东（1959-），男，教授，博导，主要研究方向为智能信息处理、系统仿真.

答案是肯定的，当前的研究中，关于机器学习流程优化的方法有贝叶斯优化、网格搜索、双层优化等。一直以来，自动机器学习(auto ML)的研究主要侧重于优化模型参数，以发现能实现分类准确度最大化的模型参数，最近的研究表明，随机搜索比穷举搜索更能有效地发现理想的模型参数[4]。尤其是利用贝叶斯优化得到的模型参数，一直有着理想的甚至优于手动调参的分类效果[6]。

本文使用基于树结构的流程优化工具(TPOT)，该工具使用遗传编程(Genetic Programming，GP)的一个模型自动地设计和优化机器学习流程[7]。本文结合雷达辐射源信号识别问题的特点与机器学习流程优化的优点，将该工具引入雷达辐射源信号识别的研究中，并通过遗传编程的相应设置在保证分类精度最大化的同时减少流程的复杂度。

1 基于树结构的自动机器学习流程

常规的机器学习，需要人根据大量实验总结的经验去构建特征、选择特征，选择合适的分类器并设置合适的分类器参数等一系列流程，对于不同的识别对象，各流程设置不尽相同，这对人工调优带来了巨大的挑战。本文使用根据遗传编程设计的基于树结构的自动机器学习流程进行流程优化，图 1 所展示的是优化后的一个典型的流程的结构，图中方框中的内容为流程中的操作算子，该结构图未注明具体的操作内容，数据集流经这些操作算子，通过对特征预处理和主成分分析得到的特征进行选择而对原始特征进行添加、删除和修改，然后选择分类模型及相应的模型参数进行分类。

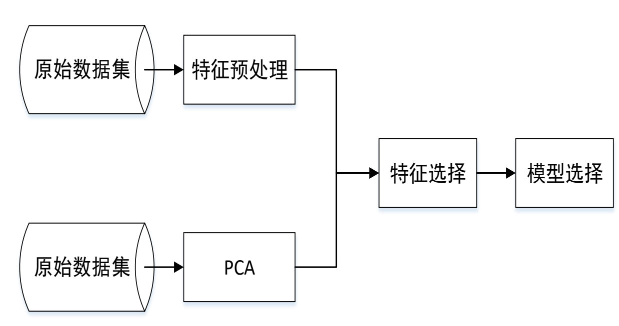


图 1 一种基于树结构的自动机器学习流程结构

1.1 操作算子

1.1.1 特征预处理

TPOT 使用的预处理方法有 Standard Scaler、Robust Scaler 和 Polynomial Features。Standard Scaler 工具根据样本的均值和方差衡量特征的重要性来对这些特征的权重进行缩放，Robust Scaler 是一个鲁棒性很强的缩放工具，采用中位数和四分位数间距去缩放特征，而 Polynomial Features 通过数值特征的多项式组合产生相互作用的特征。

1.1.2 分解

TPOT 使用 Randomized PCA 进行数据集分解降维，该方法是一种使用随机奇异值分解的主成分分析的变体。

1.1.3 特征选择

使用的特征选择方法有 RFE、Select KBest、Select Percentile 和 Select Percentile。这四种方法分别采用递归特征消除策略、选择最优 k 个特征的策略、选择最优前 n%特征的策略和移除不符合最小方差阈值的特征的策略。

1.1.4 模型选择

由于识别对象基本是有标签的数据样本，因而 TPOT 使用有监督的学习模型，包括基于树结构的决策树分类器(decision tree)、随机森林分类器(random forest)和 gradient boosting，以及支持向量机(SVM)、逻辑回归(logistic regression)和 K 近邻法(KNN)等。

以上操作算子均采用 scikit-learn 机器学习库实现，这是一个通用的 Python 机器学习库[8]。

1.2 流程的构建

上面提到的四类操作算子就是遗传编程中的树结构的节点，除此之外还有一种算子——特征合并，使用该操作算子用来将多输入合并为单输入，具体的介绍请看下面。

所有由遗传编程生成的流程，都开始于原始数据集及其副本，也就是树的叶子，在叶子之后即是四类操作算子——特征预处理、分解、特征选择和分类模型，除了这四类操作算子接在叶子之后形成的树权是随机的，操作算子的具体使用方法也是随机的。原始数据单独经过特征与处理、分解和特征选择这些操作后，输出的结果也是特征集，可以接着经过其他操作节点。若某个节点前有多个输入，则经过特征合并操作算子合并为单个特征集，再传递给下游节点。

经过模型选择节点的数据，输出结果为数据的分类预测，若不是最后的建模节点，则将该分类预测作为一种新的特征增加到输入的特征集上，并且每经过一个新的建模节点都会剔除掉原来的分类预测，另外在进化过程中，模型和其他操作的参数都是进化的对象。

最后的节点只能是建模节点或多个建模节点的合并，其预测结果作为评价该流程整体分类性能的指标。这种基于树的流程允许各节点随意变换数量和相互关系，可以实现任意结构的流程。本文将数据分为 75%的训练集和 25%的测试集，全部流程只在训练集上训练，在测试集上测试评估。

1.3 遗传编程

遗传编程的基本思想借鉴了自然界生物进化和遗传理论的原理，是一种自动随机产生搜索程序的方法。由于该算法作为一种新的全局优化搜索算法，以其简单通用、鲁棒性强，并且对非线性复杂问题显示出很强的求解能力，因而被成功地应用于许多不同的领域，并且在近几年中得到了更深入的研究[9]。

为了自动生成上节中描述的基于树结构的流程，本文使用了遗传编程的方法，具体实现方式采用一种名叫 DEAP 的 Python 库[10]。在本文的情况下，GP 构建以流程操作算子为节点的树型结构，最大限度地提高流程的最终分类精度。在这里，本文使用 GP 来演变进化作用在数据集上的操作流程以及这些

操作节点的参数，例如随机森林中的树的数量或在特征选择期间选择的特征的数量。

在本文中，GP 算法遵循标准进化算法程序，相关参数设置如表 1 所示。在每个进化过程运行开始时，系统随机生成了固定数量的树型流程，以构成遗传编程中的初代人口。然后根据其分类精度对这些流程进行评估，这里的分类精度就是这些流程个体的适应度。

表 1　遗传编程参数设置

GP 参数	内容
人口规模	100
代数	100
个体突变率	90%
个体交叉率	5%
选择方法	10%的精英保留策略，3 选 2 锦标赛选择法，再根据复杂度 2 选 1
突变	替换，插入，删除，每类突变各占 1/3
重复运行	30

在完成所有流程个体的评估后，系统进行 GP 算法的下一代生成。为了产生下一代的人口，系统首先创建了具有最高适应度的流程个体的副本，并将其置于子代的人口中，直到这些精英个体占人口总量的 10%（即 10%的精英保留策略）。为了构建下一代人口的其余部分，系统从现有人口中随机选择三个个体，然后将它们放进锦标赛中，以决定哪些个体胜出。在本次比赛中，淘汰了适应度最低的个体，然后从剩下的两个个体中选择流程复杂度更低（也就是操作节点更少）的流程，将其复制置于下一代人口中。重复这一选拔流程，直至填充完剩下的 90%的子代人口。

在下一代人口创建完毕后，系统将一个点交叉算子作用于复制后的按交叉率选择的一群个体，每次交叉将随机地选择两个个体，在流程结构中的一个随机点进行分割，然后将它们的内容彼此交换。而其余未受影响的按突变率选择的个体将进行突变：

替换突变：随机选择个体流程结构中的操作节点，将其替换为新的随机生成的流程序列。

插入突变：将一个新的随机生成的流程序列插入被插入个体的随机位置。

删除突变：随机剔除被删除个体的某段流程序列。

当某个复制后的个体被选为突变个体时，每个变异算子有 1/3 的机会作用在突变个体上。在所有交叉和变异操作中，不允许出现无效的流程，例如，将数据集传递到输入为单个参数的节点的流程是不允许产生的。

在交叉和变异操作完成以后，上一代的个体被完全删除，并以固定的代数重复这个评估-选择-交叉-变异的过程。通过这样的方式，GP 算法不断地改变流程，增加新的操作节点，提高了适应度并剔除了多余的或影响效果的操作节点。进化期间系统发现的单一性能最好的流程将被跟踪并存储在单独的位置，并在运行结束后作为流程的最终优化结果。

2　雷达辐射源信号特征集

由于笔者所参与项目的原因，本文所研究的对象为雷达辐射源信号，该数据集通过计算机仿真得到，由西南地区某电子研究所提供。数据分为两种类型，包括十部雷达脉冲截取片段及该片段的相关特征组成的特征集和脉冲时间序列。

特征集包括雷达辐射源信号的以下参数与特征：到达时间(TOA)、载频(RF)、脉宽(PW)、脉幅(PA)、到达时间差(DTOA)、脉内起始位置、脉内终止位置和调制类型（即所属雷达编号）。在该特征集中，本文选用载频、脉宽和脉幅三个特征作为原始数据集的特征集，并选用调制类型作为样本的标签。

除此之外，本文还提取了包括信息维数(IF)、信息熵(IE)、Lemple-Ziv(LZ)和小波脊频特征(Wavelet Ridge Frequency) 在内的这一系列雷达信号的脉内特征。雷达辐射源信号的脉内特征主要表现在频率、相位和幅度的变化与分布上，脉内有无调制以及采用何种调制方式，将在信号的波形上直接反映出来，因而，通过度量信号波形的复杂度可以将信号的脉内调制方式识别出来，前面提到的信息维数、信息熵和 Lemple-Ziv 都属于复杂度特征[2]。在时频平面的不同位置小波变换具有不同的分辨率，其分辨率可以针对信号的特性自动调节，这种特性使得小波变换对不同的信号具有自适应性，非常适合分析非平稳信号。小波分析既可以见到信号的全貌，又可以察看信号的细节，因而被称为“数学的显微镜”，因此可以通过小波变换特征对雷达信号调制类型进行识别[11]。

以上脉内特征皆在 MATLAB 环境下使用相应工具对脉冲时间序列进行分析提取得到，并与前面的 RF、PW 和 PA 组合在一起构成另一特征集，供后面的实验使用。

本文对得到的两个特征集中的样本进行乱序排列，每个特征集各随机选取样本 1000 个，这 1000 个样本中属于十部雷达调制类型的样本各 100 个，构成最终的两个特征集，并经过格式调整以适合优化工具的使用。

3　实验结果与分析

除了经过 TPOT 方法使用遗传编程生成的适应度最佳的机器学习流程，本文还引入了常规的一对一 SVM 方法进行对比实验，实验特征集包括上一节所述的 RF、PW 和 PA 三个特征组成的特征集 1，以及加入脉内特征的特征集 2，所有的实验都将特征集分为 75%的训练集和 25%的测试集。

图 2 和图 3 是 SVM 与 TPOT 在这两个特征集上的分类效果，显而易见，经过遗传编程得到的树型机器学习流程的分类表现更佳，从平均结果看，无论是在特征集 1 还是特征集 2 上，其分类准确率均比 SVM 高出至少 6 个百分点，并且从图中可以直观地看到其分类表现比 SVM 更稳定，而 SVM 的分类表现明显有着较大的波动。除此之外，可以看出引入脉内特征的确

能有效地提高分类准确率，不论是优化后的机器学习流程还是SVM，在加入脉内特征后分类准确率都提高了至少1个百分点。

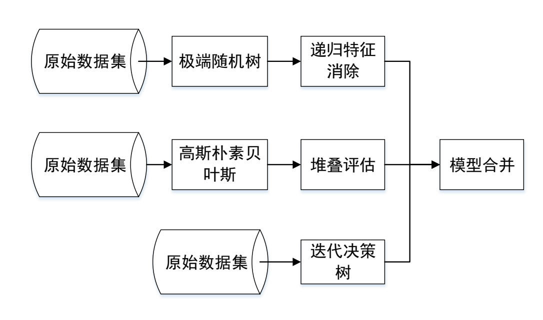


图4　TPOT在特征集2上优化得到的机器学习流程

图4展现的是在特征集2上经过遗传编程优化得到的基于树结构的机器学习流程，在这个流程中，原始数据集复制为三个副本，分别进入三个分支进行特征处理和分类，最后对这三个分类结果进行合并得到最终的分类结果，优化的结果还包括具体的参数，此处暂且不表。可以看到这其实并不是一个很简洁的流程，虽然用这样一个流程去进行分类的时间很短，但是在较大的数据集上进化出一个最终的流程可能需要较长的时间，如果模型中包含人工神经网络(ANN)的话，想必进化所耗的时间会更可观，这对于往 ANN 之类的智能学习方向发展的流程优化来说将是一个巨大的难题。

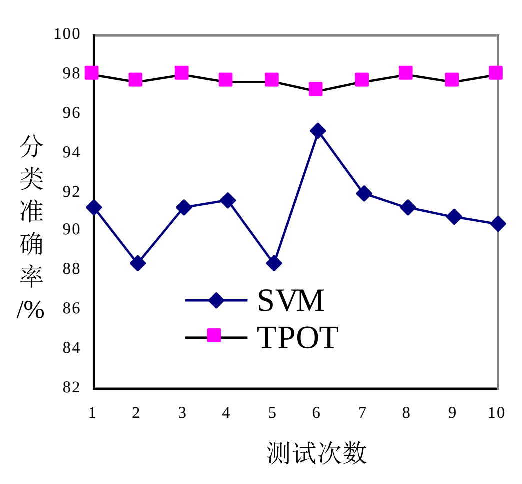


图2　SVM与TPOT在特征集1上的分类效果

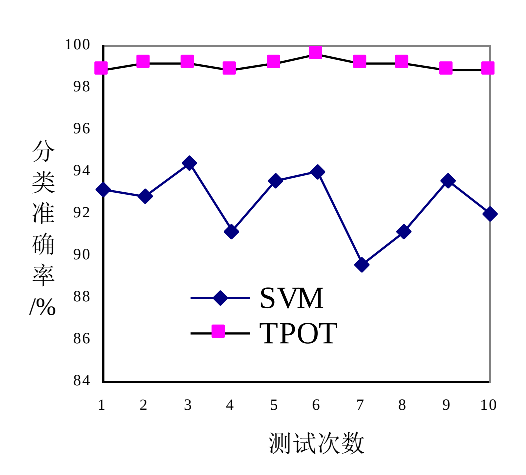


图3　SVM与TPOT在特征集2上的分类效果

4　结束语

本文从机器学习参数优化的相关研究中寻找到了另一种自动进行机器学习的方法，即自动生成树型结构的机器学习流程，并通过遗传编程优化流程的结构和相关参数，进而得到在搜索范围内最优的分类流程。

通过在雷达辐射源信号特征集上的实验，可以证明该方法能得到很好的并且较稳定的分类效果。但是如果数据集较大，该方法进行遗传操作可能需要耗费较长的时间，从本文实验的情况看，虽然最长的运算时间达到12个小时，但比起人工进行重复实验而得到效果并非很好的模型参数来说，该方法有着明显较高的效率和长远的意义。

在实验中也能看到，结合较好的特征能让自动机学习流程取得更好的结果，今后的研究中可以尝试引入效果更好的特征或特征提取算法。目前的方法并未加入 ANN 等人工智能算法，今后的研究也可以尝试为这些算法进行流程和参数优化，当然上文也提到了，这样将会耗费更为可观的时间，因此，如何在保证分类效果的同时减少进化时间将是这类研究的重点。

参考文献：

[1] 陈昌孝，何明浩，徐璟，等．雷达辐射源识别技术研究进展 [J]．空军预警学院学报, 2014, 28 (1): 1-5.

[2] 张葛祥．雷达辐射源信号智能识别方法研究 [D]．成都：西南交通大学, 2005.

[3] 闫友彪，陈元琰．机器学习的主要策略综述 [J]．计算机应用研究, 2004, 21 (7): 4-10.

[4] Olson R S, Bartley N, Urbanowicz R J, et al. Evaluation of a tree-based pipeline optimization tool for automating data science [C]// Proc of Genetic and Evolutionary Computation Conference. 2016: 485-492.

[5] Olson R S, Moore J H. TPOT: A tree-based pipeline optimization tool for automating machine learning [J]. Journal of Machine Learning Research, 2016, 64: 66-74.

[6] Klein A, Falkner S, Bartels S, et al. Fast Bayesian optimization of machine learning hyperparameters on large datasets [J/OL]. 2016. https://arxiv.org/abs/1605. 07079.

[7] 李少波，胡建军．遗传编程与机电系统创新设计 [M]．北京：机械工业出版社, 2009: 61-62.

[8] Pedregosa F, Varoquaux G, Gramfort A, Michel V, et al. Scikit-learn: machine learning in python [J]. J Mach. Learn. Res, 2011, 12: 2825-2830.

[9] 查志琴，高波，郑成增．遗传编程实现的研究 [J]．计算机应用, 2003, 23 (7): 137-139.

[10] Fortin, F. A, Gardner, M. A, Parizeau, M, Gagne, C, et al. DEAP: evolutionary algorithms made easy [J]. J Mach. Learn. Res, 2012, 13: 2171-2175.

[11] 余志斌．基于脉内特征的雷达辐射源信号识别研究 [D]．成都：西南交通大学, 2010.